DOCUMENT RESUME

ED 447 210 TM 032 118

AUTHOR Henson, Robin K.
TITLE A Primer on Coefficient Alpha.
PUB DATE 2000-11-16
NOTE 41p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Bowling Green, KY, November 15-17, 2000).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Reliability; *Scores; Test Results; Test Theory
IDENTIFIERS *Alpha Coefficient

ABSTRACT

Because reliability is a function of scores, and not tests per se, it is inaccurate to hold that a given test will yield scores with the same reliability across samples. Therefore, score reliability should always be reported and interpreted in both measurement and substantive studies. In an effort to facilitate this outcome, this paper is intended to provide an interpretive framework for applied researchers and others seeking a conceptual understanding of score reliability. The paper reviews some basic tenets of classical test theory, discusses the salient factors that affect reliability estimates, with emphasis on coefficient alpha, and present several suggestions toward a better understanding, and improved use, of score reliability. (Contains 6 tables, 1 figure, and 43 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 447 210

Running head: Coefficient alpha

A Primer on Coefficient alpha

Robin K. Henson

University of North Texas 76203-1337

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Robin Henson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY, November, 16, 2000. Correspondence concerning this manuscript should be sent to rhenson@tac.coe.unt.edu.

TM032118

BEST COPY AVAILABLE

Abstract

Because reliability is a function of scores, and not tests per se, it is inaccurate to hold that a given test will yield scores with the same reliability across samples. Therefore, score reliability should always be reported and interpreted in both measurement and substantive studies. In an effort to facilitate this outcome, the present paper is intended to provide an interpretive framework for applied researchers and others seeking a conceptual understanding of score reliability. The paper will: a) review some basic tenets of classical test theory, b) discuss the salient factors that affect reliability estimates, with emphasis on coefficient alpha, and c) present several suggestions toward a better understanding (and use) of score reliability.

A Conceptual Primer on Coefficient alpha

In a recently published (and important) report, the American Psychological Association (APA) Task Force on Statistical Inference declared the need for all studies to report measures of effect size along with their statistical significance results (Wilkinson & APA Task Force on Statistical Inference, 1999). The Task Force noted:

> It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p-value or, better still, a confidence interval. . . . Always provide some effect-size estimate when reporting a p-value. (p. 599, emphasis added)

The Task Force went on to state, "Always present effect sizes for primary outcomes. . . . It helps to add brief comments that place these effect sizes in a practical and theoretical context" (p. 599, emphasis added).

The mandate to "always" report effect sizes is an important step beyond the fourth edition of the APA's Publication Manual, which only recommended reporting of effect sizes in research (APA, 1994, p. 18). Empirical studies, however, have shown that this recommendation has had little impact on researchers' inclusion of effect size information in their articles and even less impact on consultation of effects for "practical and

theoretical context" (cf. Henson & Smith, 2000; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000).

Furthermore, the Task Force (Wilkinson & APA Task Force on Statistical Inference, 1999) also recommended that authors "provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (p. 596). This recommendation to report score reliability in all studies relates directly to the mandate to also include effect sizes, because "Interpreting the size of observed effects requires an assessment of the reliability of the scores" (p. 596). Effect size magnitude is inherently attenuated by the reliability of the scores used to obtain the effect estimate (Reinhardt, 1996). As Reinhardt (1996) observed,

> Reliability is critical in detecting effects in substantive research. For example, if a dependent variable is measured such that the scores are perfectly unreliable, the effect size in the study will unavoidably be zero, and the results will not be statistically significant at any sample size, including an incredibly large one. (p. 3)

As a point of illustration, the maximum r^2 between two variables equals the product of the square root of the reliabilities (cf. Locke, Spirduso, & Silverman, 1987, p. 28), such that when one variable has alpha = .70 and another variable

has alpha = .60, the maximum possible effect would be [(.70)**.5][(.60)**.5] = (.8367)(.7746) = .6481 = r^2. Accordingly, the reliability of the scores in any study, measurement and substantive, is central to understanding the observed relationships between variables. Because all classical analyses (e.g., t-test, ANOVA, regression) are part of the same general linear model and are correlational in nature (Bagozzi, Fornell & Larcker, 1981; Cohen, 1968; Henson, 2000; Knapp, 1978; Thompson, 1991), most studies should report and interpret results in light of reliability estimates (Thompson, 1994).

Unfortunately, too few researchers report score reliability for their studies and even fewer interpret their effects in light of reliability. This deficit in the literature is likely due to myriad factors, the chief of which is the common misconception that reliability inures to tests, rather than scores (cf. Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998). A contrary view is given by Sawilowsky (2000a, 2000b).

Indeed, it is scores, not tests, that are either reliable or unreliable. Furthermore, a given test may yield grossly divergent score reliability estimates upon different administrations. The reader is referred to Caruso (2000); Henson, Kogan, and Vacha-Haase (in press); Viswesvaran and Ones (2000); Yin and Fan (2000), and Vacha-Haase (1998) for examples of this phenomenon.

Because reliability inures to scores, different samples, testing conditions, and any other factor that may impact observed scores can in turn affect reliability estimates. Because score reliability inherently attenuates effect sizes, it also will impact statistical power, an often overlooked point (Onwuegbuzie & Daniel, 2000). Because effects and power may be attenuated by the reliability of observed scores, reliability should always be reported and considered in result interpretation (Wilkinson & APA Task Force on Statistical Inference, 1999).

Purpose

Pedhazur and Schmelkin (1991) suggested that many researchers' misconceptions and unawareness surrounding score reliability may be due to decreased emphasis on measurement coursework in doctoral programs. Aiken et al. (1990) verified this measurement vacuum in doctoral curricula. In a national survey of American Educational Research Association (AERA) members, Mittag and Thompson (2000) found less than desirable understanding of score reliability among respondents. While reliability is relevant for most situations, the issue is particularly salient in applied studies, where previously developed measures are often used to answer substantive research questions. In these cases, it is the reliability of the presently obtained scores, not the reliabilities reported from

test manuals or previous studies, that bears directly on substantive interpretations.

Accordingly, the present paper is intended to provide an interpretive framework for applied researchers and others seeking a conceptual understanding of score reliability. The paper will: a) review some basic tenets of classical test theory, b) discuss the salient factors that affect reliability estimates, with emphasis on coefficient alpha, and c) present several suggestions toward a better understanding (and use) of score reliability.

Some Basic Tenets of Classical Test Theory

Reliability is concerned with score accuracy. Obviously, it is important that our scores are accurate, particularly when there are important ramifications of our interpretations. The more measurement error that exists in our scores, the less useful these scores may be for analysis and interpretation. This section addresses several key points related to the classical test theory underlying many reliability estimates. The reader is referred to Crocker and Algina (1986) for a complete treatment.

Ratio of Score Variances: The General Linear Model in Measurement

The classical conceptualization of score reliability relates the concept of score accuracy to "true scores." In

other words, for any measurement occasion that is less than perfect, a set of scores will contain variance that is true score variance (accurately measuring the trait of interest) and variance that is due to error (factors inhibiting accurate measurement, e.g., fatigue, confusing questions). The sum of these two variances yields the total score variance of the observed scores, such that:

$$\sigma_{TRUE}^2 + \sigma_{ERROR}^2 = \sigma_{TOTAL}^2$$

Graphically, an example of this relationship may be depicted by Figure 1. Here only 80% of the total score variance is attributable to true (accurate) score variance and the remaining 20% is attributable to error. In this case, the coefficient alpha would be .80 (this statistic will be discussed in detail later), indicating that 80% of the total score variance is reliable.

INSERT FIGURE 1 ABOUT HERE

Figure 1 makes explicit the reason effect sizes are inherently attenuated by reliability. Only reliable variance may be correlated between any two variables (or linear composite sets of variables beyond the bivariate case). It is impossible to correlate random error across variables, thereby attenuating an r^2 type effect size to be less than 1.00.

Another generalization of Figure 1 informs us that score reliability can be conceptualized as a ratio of true score variance to total (observed) score variance (80% in Figure 1). Dawson (1999) noted that coefficient alpha was an analog of the more familiar $\underline{r}^2$ type effect, and accordingly represents a ratio of variances. Dawson generalized the $\underline{r}^2$ statistic and noted:

> One alternative formula with which to compute the $\underline{r}^2$ effect size is:
>
> $$\underline{r}^2 = SOS_{EXPLAINED} / SOS_{TOTAL}. \quad (1)$$
>
> . . . Formula (1) is a general formula for effect for all parametric univariate methods. For example, this formula is correct for $\underline{r}^2$, for $\underline{R}^2$ (a regression effect size), and eta^2 (an ANOVA and $\underline{t}$-test effect size). Conceptually, this formula asks, "what portion (or percentage) of the total information can an extraneous variable explain or predict?" Thus, any variance-accounted-for r^2 effect size is a ratio of variances; the formula could also be written as:
>
> $$\underline{r}^2 = V_{EXPLAINED} / V_{TOTAL} \quad (2)$$
> $$= [SOS_{EXPLAINED} / (n - 1)] / SOS_{TOTAL} / (n - 1).$$
>
> Because formula (2) contains n-1 in both the numerator and the denominator, and these terms cancel, formula (1) is the more usual and convenient expression of this very general formula. (pp. 105-106)

For coefficient alpha (a common estimate of reliability; Cronbach, 1951), this same ratio of variances is apparent in the formula:

$$\alpha = \underline{k} \,/(\underline{k} - 1)\, [1 - (\Sigma\sigma_k^2 \,/\, \sigma_{TOTAL}^2)],$$

where $\underline{k}$ = the number of items on the test, $\Sigma\sigma_k^2$ = the sum of all the $\underline{k}$ item variances, and σ_{TOTAL}^2 = the variance of the total test scores. In the alpha formula, the ratio of variances is captured in the $(\Sigma\sigma_k^2 \,/\, \sigma_{TOTAL}^2)$ term.

Because of this ratio of variances, Dawson (1999) noted that the general linear model which guides much substantive statistical analysis also infuses the measurement context: "The presence of the general linear model (GLM) across both substantive and measurement analyses can also be seen in the computation of coefficient alpha (Cronbach, 1951) as the ratio of two variances" (p. 109). However, as Thompson noted (1999), "psychometrically alpha involves more than only variances and their ratios to each other" (p. 12). Most explicitly, the alpha formula invokes $\Sigma\sigma_k^2$ as the numerator, which is related to, but different from, the $SOS_{EXPLAINED}$ noted above (this issue will be explained momentarily along with illustration of coefficient alpha).

Estimates of Measurement Error

Typically, many authors conceptualize three sources of measurement error within the classical framework: content sampling of items, stability across time, and interrater error (see e.g., Anastasi & Urbina, 1997; Hopkins, 1998; Popham, 2000). Content sampling refers to the theoretical idea that the test is made up of a random sampling of all possible items that could be on the test. If so, the items should be highly interrelated, theoretically because they assess the same construct of interest (e.g., self-esteem, achievement). This item interrelationship is typically called internal consistency, which suggests that the items on a measure should correlate highly with each other if they truly represent appropriate content sampling. If items are highly correlated, it is theoretically assumed that the construct of interest has been measured to some degree of accuracy (i.e., the scores are reliable).

As a measure of internal consistency and a generalization of the older split-half method, Kuder and Richardson (1937) presented their classic formula, KR-20 (named such because the formula was the 20th listed in their article), as:

$$KR\text{-}20 = \underline{k} \,/\, (\underline{k} - 1)\, [1 - (\Sigma p_k q_k \,/\, \sigma_{TOTAL}^2)],$$

where $\underline{k}$ = the number of items on the test, p_k = the proportion of people answering item $\underline{k}$ correctly, q_k = the proportion of people

answering item k incorrectly (i.e., $1 - p_k$), and σ_{TOTAL}^2 = the variance of the total test scores. Because $\Sigma p_k q_k$ deals with mutually exclusive proportions for two possible outcomes, it should be clear that KR-20 only works when test items are dichotomously scored (e.g., 0 and 1). This formula may apply to either achievement or attitude measures, as long as scoring is dichotomous (e.g., correct v. incorrect, agree v. disagree).

Importantly, the variance of a dichotomously scored item (σ_k^2) will equal $p_k q_k$, always. If all persons responded the same to an item, then $\sigma_k^2 = p_k q_k = 0$, because no variance would be present in the scores. Furthermore, if one-half of the responses were scored "0" and the other half scored "1", then the scores would have maximum variability. When items are dichotomously scored, the maximum variability possible is $\sigma_k^2 = p_k q_k = .25$. This is because each squared deviation score will be .25, a result of subtracting the mean of .5 from 0 or 1 and squaring this difference. The sum of these squared deviation scores (i.e., sum of squares) divided by n (variance) will result in .25, regardless of sample size (cf. Reinhardt, 1996).

Fourteen years after the advent of KR-20, Cronbach (1951) introduced coefficient alpha, a more general form of the KR-20 formula. With specific terms defined above, coefficient alpha is given as:

$$\alpha = \underline{k} / (\underline{k} - 1) \; [1 - (\Sigma\sigma_k^2 / \sigma_{TOTAL}^2)],$$

Comparison of the KR-20 and alpha formulae reveals that only the numerator of the variance ratio differs. Because $\Sigma\sigma_k^2 = \Sigma\underline{p}_k\underline{q}_k$ as noted above, it should be apparent that alpha can be used with dichotomously scored items. However, because the sum of the item variances is used as the numerator (and not $\Sigma\underline{p}_k\underline{q}_k$ per se), alpha can also be used with measures employing multiple response categories such as Likert scale data.

In both KR-20 and alpha, it is clear that certain data features will lead to higher reliability estimates. Holding the number of items constant ($\underline{k}$), reliability will increase as the sum of item variances decreases and the total score variance increases.

A second source of measurement error involves the occasion of measurement. Often, a test-retest reliability estimate (correlation between scores on two occasions by the same sample) is calculated to evaluate score stability. If we have accurately measured someone on the trait if interest with a test, we should be able to accurately measure them again later. The degree that our two sets of scores do not correlate indicates measurement error due to time of measurement. Here a fundamental tenet of classical test theory is illustrated. As explained by Henson et al. (in press):

> In terms of classical measurement theory (holding the number of items on the test and the sum of item variances constant), increased variability of total scores suggests that we can more reliably order people on the trait of interest, and thus more accurately measure them. This assumption is made explicit in the test-retest reliability case, when consistent ordering of people across time on the trait of interest is critical in obtaining high reliability estimates.

If the ordering of subjects changes from one testing occasion to the other, then certainly our accuracy (reliability) in measuring them is less than perfect. Accordingly, classical reliability estimates hinge on the variance of the total scores. As this variance increases, the reliability estimate will also tend to increase, due to greater theoretical confidence that we have accurately ordered (measured) the subjects on the trait of interest.

One implication of this role of total score variance is that different samples will likely yield different score reliabilities because the total variance will likely change. For example, Thompson (1994) observed: "The same measure, when administered to more heterogeneous or more homogeneous sets of subjects, will yield scores with differing reliability" (p. 839).

A third source of measurement error, interrater variation, is only applicable when scores are derived from raters. Because most

testing situations do not involve raters, this source will not be discussed here.

Importantly, these sources of measurement error are separate and cumulative (Anastasi & Urbina, 1997). Too many researchers believe that if they obtain alpha = .90 for their scores, then the same 10% of measurement error would be found in a test-retest or interrater coefficient. Instead, assuming 10% error for internal consistency, stability, and interrater, then the overall meaurement error would be 30%, not 10%, as these estimates explain different sources of error. As an aside, generalizability theory (as opposed to classical test theory) allows for the simultaneous examination of these sources of error as well as the interactions between them using ANOVA methodology. The interested reader is referred to Kieffer (1999) and Shavelson and Webb (1991) for accessible treatments of G theory.

A Conceptual Primer on Coefficient alpha

As noted, alpha invokes a general linear model ratio of explained variance to total variance as a fundamental component in its calculation. However, as a measure of internal consistency, it also must account for the intercorrelation among the items, with the assumption that as items are more highly correlated, the magnitude of alpha will increase.

Three heuristic examples are used here to illustrate the salient data features that impact coefficient alpha. These

examples are heavily dependent on Thompson (1999) and Reinhardt (1996) and are adapted for use here.

Example One: Perfectly Uncorrelated Items

Although test items often are correlated to some degree, the present example illustrates the impact on alpha when items are perfectly uncorrelated (r and covariance = 0 for all pairwise item combinations). Table 1 presents a heuristic data set for four test items with inter-item correlations of 0. [Note as well that $r_{XY} = COV_{XY} / \{(SD_X)(SD_Y)\}$, and also $COV_{XY} = r_{XY} \{(SD_X)(SD_Y)\}$.]

INSERT TABLES 1 AND 2 ABOUT HERE

Based on the above formula for alpha, reliability can be computed if we can identify the number of items, the sum of the item variances, and the variance of the total scores. The first two of these items is given by Table 1, with k = 4 and the sum of the item variances as .73 = (.22 + .18 + .18 + .15). Crocker and Algina (1986, p. 95) presented a formula for the calculation of the total score variance using only the Table 1 data:

$$\sigma_{TOTAL}^2 = \Sigma\sigma_k^2 + [\Sigma COV_{ij} \text{ (for } i<j) * 2]$$

Close examination of this formula reveals that total test score variance can be conceptualized as an additive function of two components: a) the sum of the item variances ($\Sigma\sigma_k^2$) and b) the doubled sum of the unique covariances [ΣCOV_{ij} (for $i<j$) * 2].

This formula highlights the important point that the total test score variance is at least partially dependent on the intercorrelations among the items on a test, a finding in harmony with the idea that alpha is a measure of internal consistency. Table 2 presents calculations for determining the covariance portion of the total test score variance. Table 2 also illustrates the COV to r transformation as noted above. Using the data from Tables 1 and 2, the total test score variance is found with:

$$\begin{aligned}\sigma_{TOTAL}^2 &= \Sigma\sigma_k^2 + [\Sigma COV_{ij} \text{ (for } i<j) * 2] \\ &= (.22 + .18 + .18 + .15) + .00 \\ &= .73.\end{aligned}$$

These calculations indicate that in this example the total score variance is only a function of the sum of the individual item variances, because the covariances were 0. This finding verifies that "only when the covariances among items are 0 will SD^2 [i.e., total score variance] equal Σpq" (Sax, 1974, p. 182).

Now using the total score variance as our last remaining piece of information, alpha can be found with:

$$\begin{aligned}\alpha &= \underline{k} \,/(\underline{k} - 1)\ [1 - (\Sigma\sigma_k^2 \,/\, \sigma_{TOTAL}^2)], \\ &= 4 \,/(4 - 1)\ [1 - (.22 + .18 + .18 + .15) \,/\, .73] \\ &= 4 \,/\, 3 \qquad [1 - (.73 \,/\, .73)] \\ &= 1.33 \qquad [1 - 1] \\ &= 1.33 \qquad [0] \\ &= 0.\end{aligned}$$

Because the items shared no variance, such that the covariances and correlations were 0, it stands to reason that

there was no internal consistency among the items. Accordingly, alpha's calculations led to this logical conclusion ($\alpha = 0$). Furthermore, based on this understanding, the alpha formula reveals that we should expect alpha to increase as the covariances contribute more to the total score variance.

Example Two: Perfectly Correlated Items

When items are perfectly correlated, and thereby possessing perfect internal consistency, we should no doubt expect alpha to reach its maximum of 1 (representing 100% of true score variance due to content sampling). Table 3 presents data on four perfectly correlated test items. Table 4 presents the calculations necessary to obtain the total score variance using the Crocker and Algina (1986, p. 95) formula. Using these results, the total score variance is:

$$\begin{aligned}\sigma_{TOTAL}^2 &= \Sigma\sigma_k^2 + [\Sigma COV_{ij} \text{ (for } i<j) * 2] \\ &= (.22 + .18 + .18 + .15) + (1.08 * 2) \\ &= .73 + 2.16 \\ &= 2.89.\end{aligned}$$

Using the total score variance, alpha is:

$$\begin{aligned}\alpha &= \underline{k} /(\underline{k} - 1) \; [1 - (\Sigma\sigma_k^2 / \sigma_{TOTAL}^2)] \\ &= 4 /(4 - 1) \; [1 - (.22 + .18 + .18 + .15) / 2.89] \\ &= 4 / 3 \qquad [1 - .73 / 2.89] \\ &= 1.33 \qquad [1 - .2525952] \\ &= 1.33 \qquad [.7474048] \\ &= .9940.\end{aligned}$$

As expected, alpha = 1 (within rounding error due to calculation of the covariances in Table 3), indicating perfect internal consistency of scores.

INSERT TABLES 3 AND 4 ABOUT HERE

Example Three: Perfectly Correlated Items with Mixed Signs

It is possible for items to be highly correlated but not all in the same direction. Table 5 presents the heuristic data matrices for perfectly correlated items with but with mixed signs and Table 6 presents calculations that lead to the total score variance. The total score variance is:

$$\begin{aligned}\sigma_{TOTAL}^2 &= \Sigma\sigma_k^2 + [\Sigma COV_{ij} \text{ (for } i<j) * 2] \\ &= (.22 + .18 + .18 + .15) + (-.08 * 2) \\ &= .73 + (-.16) \\ &= .57.\end{aligned}$$

Coefficient alpha is solved as:

$$\begin{aligned}\alpha &= \underline{k} /(\underline{k} - 1) \; [1 - (\Sigma\sigma_k^2 / \sigma_{TOTAL}^2)] \\ &= 4 /(4 - 1) \; [1 - (.22 + .18 + .18 + .15) / .57] \\ &= 4 / 3 \qquad [1 - .73 / .57] \\ &= 1.33 \qquad [1 - 1.2807018] \\ &= 1.33 \qquad [-.2807018] \\ &= -.3733.\end{aligned}$$

Here we have found what Thompson (1999, p. 15) called a "paradox" in the calculation of alpha. That is, how can alpha be negative, given that it is a squared metric statistic ($\underline{r}^2$ type ratio of variances)! Solving for alpha with the equivalent formula presented by Sax (1974, p. 181) helps provide a deeper understanding of alpha's ratio of variances:

$$\begin{aligned}\alpha &= \underline{k} /(\underline{k} - 1) \; [(\sigma_{TOTAL}^2 - \Sigma\sigma_k^2) / \sigma_{TOTAL}^2)] \\ &= 4 /(4 - 1) \; [(.57 - .73) / .57] \\ &= 4 / 3 \qquad [-.16 / .57] \\ &= 1.33 \qquad [-.2807018] \\ &= -.3733.\end{aligned}$$

Here we find that the numerator essentially represents the covariances between the test items. This follows from the Crocker and Algina (1986, p. 95) formula used to calculate the total score variance, which shows total score variance as an additive function of the sum of the item variances and the doubled sum of the unique item covariances:

$$\sigma_{TOTAL}^2 = \Sigma\sigma_k^2 + [\Sigma COV_{ij} \text{ (for } i<j) * 2].$$

In the numerator of the alpha formula above, we have essentially removed the sum of the item variances ($\Sigma\sigma_k^2$) from the total score variance (σ_{TOTAL}^2), which leaves the summed item covariances [ΣCOV_{ij} (for $i<j$) * 2]. The covariance term is found in the bolded calculations for alpha above (**-.16**) and in the calculations in Table 6. Thus, the alpha ratio includes the sum of the item covariances over the total score variance. Inspection of the Crocker and Algina (1987, p. 95) formula for total score variance reveals that we would expect alpha to increase when the item correlations are large and in the same direction. This ratio of a "covariance" to a "variance" is legitimate. As Thompson (1999) explained:

> Is the ratio of the sum of item score covariances to the total score variance a ratio of apples to oranges (i.e., of two unlike entities to each other)? No, because, in addition to both being in a squared metric . . . the [total

score variance] . . . is itself in part a function of covariances. . . (p. 15-16)

The negative result (α = -.37) we find in the present example, then, is a mathematical artifact that occurs when the sum of the item variances exceeds the total score variance. Conceptually, this would mean that the individual variability of the k items tends to be greater than the shared variability (covariance/correlation) between the k items. If this is true, then internal consistency suffers because the items appear to be measuring different constructs! In keeping with a classical test theory perspective, the psychometric properties of alpha (and KR-20) capture this conceptual expectation.

Toward a Better Understanding (and Use) of Score Reliability

As noted, many researchers fail to report score reliability for their data, leaving the reader to guess whether the scores were reliably measured and to what degree, if any, the observed effects were attenuated by measurement error. Furthermore, it is all too common to see researchers referring to the "reliability of the test" when, in fact, reliability inures to scores, not tests, and can vary considerably across samples.

The etiology of these errors in reporting practice is likely complex. However, as Thompson and Vacha-Haase (2000) noted, "some people use the phrase 'the reliability of the test' as a telegraphic shorthand in place of truthful but longer

statements (e.g., 'the reliability of the test scores')" (p. 178). Worthen, White, Fan, and Sudweeks (1999) also noted: "many have adopted the shorthand of speaking of the test's reliability, a sin that can probably be forgiven as long as you understand this critical distinction [between reliability of scores versus tests]" (p. 95, emphasis in original). Unfortunately, as Thompson (1992) explained, "the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice" (p. 436).

Pedhazur and Schmelkin (1991) placed a portion of the blame on inadequate doctoral curricula, noting that

> although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. Thus, many students get the impression that no special competencies are necessary for the development and use of measures. (p. 2-3)

In an empirical evaluation of doctoral curricula, Aiken et al. (1990) also noted little emphasis on measurement issues. With the above discussion in mind, the following items are presented in effort to help further better understanding and use of score reliability.

Understand that Reliability Affects Power

Reliability inherently attenuates the maximum possible magnitude of relationships between variables (see above for discussion of attenuation of effect size). Accordingly, all else being constant, poor score reliability will reduce the power of statistical significance tests (cf. Onwuegbuzie & Daniel, 2000). When effects are reduced, they become harder to find. Researchers would be compelled to increase sample size or their $p_{CRITICAL}$ level to compensate for this loss of power.

When researchers find non-statistically significant results due to poor measurement, the bottom-line ramifications may include greater difficulty publishing in a literature biased toward statistically significant results, ignoring potentially meaningful effects, and a perpetuated misunderstanding of why the results were not statistically significant (i.e., ignoring a potential measurement problem). For a more complete discussion of statistical significance tests, the reader is referred to the seminal work of Cohen (1990, 1994) as well as Henson and Smith (2000) and Thompson (1994, 1996).

Reporting Practices and Interpretation

Researchers should report reliability for the scores at hand, and not depend on estimates from prior studies or test manuals. As correctly noted by Gronlund and Linn (1990), "Reliability refers to the results obtained with an evaluation

instrument and not to the instrument itself. Thus it is more appropriate to speak of the reliability of 'test scores' or the 'measurement' than of the 'test' or the 'instrument'" (p. 78, emphasis in original). Furthermore, researchers would do well to use precise language when referencing the reliability of their scores.

Unfortunately, empirical studies confirm that very few researchers actually report reliability estimates for their data (cf. Caruso, 2000; Vacha-Haase, 1998; Yin & Fan, 2000). For example, Yin and Fan observed that only 7.5% of articles employing the Beck Depression Inventory reported precise reliability estimates for the data in hand. Examples of inaccurate language use are also common.

Because reliability affects power by attenuating effect sizes, results should be interpreted in light of the obtained reliability. Small effects may be due, in part, to poor measurement. Furthermore, large effects are only possible to the degree allowed by the integrity of the scores. Outcomes on statistical significance tests may be adversely affected by measurement problems. Unfortunately, because so few researchers report reliability, and even fewer interpret results in light of reliability, the impact of this phenomenon is unknown. As researchers report reliability for the data in hand, and consider these estimates when interpreting their results, more will be

learned about reliability's impact on power, effect sizes, and statistical signficance tests.

Reliability Generalization Studies

Because reliability may, and does, vary upon different administrations of a test, Vacha-Haase (1998) employed a meta-analytic method called "reliability generalization" (RG) that allows examination of the variability of score reliability across studies. In addition, coded study characteristics (such as composition and variability) can be used as potential predictors of reliability variation, thereby providing some evidence of which sampling conditions most impact score reliability. Vacha-Haase's method is based on the older validity generalization approach (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977), and represents an important development in the examination of score integrity.

A primary benefit of RG studies is the cumulative information they may yield in describing study characteristics that impact reliability estimates for scores from a given test, and, perhaps, study characteristics that consistently impact score reliability across different tests. It is also possible to characterize score reliability for constructs, rather than for scores on a single test per se. For example, Henson et al. (in press) examined the construct of teacher self-efficacy across several instruments.

In order for the benefit of RG studies to be realized, however, multiple RG studies must be conducted and receive recognition in the published literature. One significant barrier to this benefit is the failure of researchers to report reliability coefficients for the scores at hand (which become the dependent variable in an RG study). As metaphorically illustrated by Thompson and Vacha-Haase (2000),

> . . . it is important to remember that RG studies are a meta-analytic characterization of what is hoped is a population of previous reports. We may not like the ingredients that go into making this sausage, but the RG chef can only work with the ingredients provided by the literature. (p. 184)

Accordingly, reporting of reliability coefficients would not only inform the study in which the reliability was reported, but also facilitate meta-analytic RG studies. Readers are referred to Vacha-Haase (1998) and Thompson and Vacha-Haase (2000) for more compete discussions of RG.

Summary

From a classical test theory perspective, score reliability relates to true score variance in a set of observed scores. It is presumed that the true score variance represents an accurate measurement of the construct of interest. There are a variety of classical test theory reliability estimates, including

internal consistency and test-retest coefficients. The present paper presented a conceptual understanding of a measure internal consistency, coefficient alpha, as an index of the ratio of true to total score variance. Importantly, reliability is a function of the scores obtained for a given measure, and are not a function of the measure/test itself. Therefore, researchers ought to report reliability for the data at hand and interpret results in light of the obtained estimates. This practice would move the field toward a better understanding and use of score reliability. It would also facilitate more (and more accurate) reliability generalization studies that characterize measurement error across test administrations.

References

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., with Roediger, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). The training in statistics, methodology, and measurement in psychology. American Psychologist, 45, 721-734.

American Psychological Association (1994). Publication Manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Bagozzi, R.P., Fornell, C., & Larcker, D.F. (1981). Canonical correlation analysis as a special case of a structural relations model. Multivariate Behavioral Research, 16, 437-454.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. Educational and Psychological Measurement, 60, 236-254.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Chicago: Holt, Rinehardt and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 197-334.

Dawson, T. E. (1999). Relating variance portioning in measurement analyses to the exact same process in substantive analyses. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 101-110). Stamford, CT: JAI Press.

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Henson, R. K. (2000). Demystifying parametric analyses: Illustrating canonical correlation as the multivariate general linear model. Multiple Linear Regression Viewpoints, 26(1), 11-19.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (in press). A reliability generalization study of the Teacher Efficacy Scale and related instruments. Educational and Psychological Measurement.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force report and current trends. Journal of Research and Development in Education, 33, 285-296.

Hopkins, K. D. (1998). Educational and psychological measurement and evaluation (8th ed.). Boston: Allyn and Bacon.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis. Newbury Park, CA: Sage.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (1987). Proposals that work: A guide for planning dissertations and grant proposals (2nd ed.) Newbury Park, CA: Sage.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. Educational Researcher, 29(4), 14-20

Onwuegbuzie, A. J. & Daniel, L. G. (2000, November). Reliability generalization: The important of considering sample specificity, confidence intervals, and subgroup differences. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum.

Popham, W. J. (2000). Modern educational measurement: Practical guidelines for educational leaders (3rd ed.). Boston: Allyn and Bacon.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in

social science methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Sawilowsky, S. S. (2000a). Reliability: Rejoinder to Thompson and Vacha-Haase. Educational and Psychological Measurement, 60, 196-200.

Sawilowsky, S. S. (2000b). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. Educational and Psychological Measurement, 60, 157-173.

Sax, G. (1974). Principles of educational measurement and evaluation. Belmont, CA: Wadsworth.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Shavelson, R., & Webb, N. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24, 80-95.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26-30.

Thompson, B. (1999, February). Understanding Coefficient alpha, Really. Paper presented at the annual meeting of the Educational Research Exchange, College Station, TX.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement, 60, 174-195.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. Theory & Psychology, 10, 413-425.

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. Educational and Psychological Measurement, 60, 224-235.

Wilkinson, L., & American Psychological Association (APA) Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American

Psychologist, 54, 594-604. (Reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html)

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). Measurement and assessment in schools. New York: Addison Wesley Longman.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. Educational and Psychological Measurement, 60, 201-223.

Table 1

Example One: Item Correlations (Covariances) Are 0

	Correlation				Variance/Covariance			
Var.	1	2	3	4	1	2	3	4
1	1.00				.22			
2	.00	1.00			.00	.18		
3	.00	.00	1.00		.00	.00	.18	
4	.00	.00	.00	1.00	.00	.00	.00	.15

Note. Item score variances are underlined and represent the diagonal of the variance/covariance matrix.

Table 2

Calculation of Total Test Score Variance (σ_{TOTAL}^2) for Example One

Pairing		COV/Variance			r/SD			
i	< j	COV_{ij}	σ_i^2	σ_j^2	r_{ij}	SD_i	SD_j	COV_{ij}'
1	2	.00	.22	.18	.00	.47	.42	.00
1	3	.00	.22	.18	.00	.47	.42	.00
1	4	.00	.22	.15	.00	.47	.39	.00
2	3	.00	.18	.18	.00	.42	.42	.00
2	4	.00	.18	.15	.00	.42	.39	.00
3	4	.00	.18	.15	.00	.42	.39	.00

ΣCOV_{ij} = .00

$\Sigma COV_{ij} * 2$ = .00

Note. COV_{ij}' represents the recalculated covariance using $COV_{ij}' = r_{ij} (SD_i * SD_j)$. These estimates match the original covariances (COV_{ij}) and illustrate the r to COV transformation.

Table 3

Example Two: Item Correlations (Covariances) Are 1

	Correlation				Variance/Covariance			
Var.	1	2	3	4	1	2	3	4
1	1.00				.22			
2	1.00	1.00			.20	.18		
3	1.00	1.00	1.00		.20	.18	.18	
4	1.00	1.00	1.00	1.00	.18	.16	.16	.15

Note. Covariances were found with $COV_{ij} = r_{ij} (SD_i * SD_j)$, where the standard deviations are the square root of the variance for the variable. Covariances are rounded to two decimal places.

Table 4

Calculation of Total Test Score Variance (σ_{TOTAL}^2) for Example Two

Pairing		COV/Variance			r/SD			
i	< j	COV_{ij}	σ_i^2	σ_j^2	r_{ij}	SD_i	SD_j	COV_{ij}'
1	2	.20	.22	.18	1.00	.47	.42	.20
1	3	.20	.22	.18	1.00	.47	.42	.20
1	4	.18	.22	.15	1.00	.47	.39	.18
2	3	.18	.18	.18	1.00	.42	.42	.18
2	4	.16	.18	.15	1.00	.42	.39	.16
3	4	.16	.18	.15	1.00	.42	.39	.16

ΣCOV_{ij} = 1.08
$\Sigma COV_{ij} * 2$ = 2.16

Note. COV_{ij}' represents the recalculated covariance using $COV_{ij}' = r_{ij} (SD_i * SD_j)$. These estimates match the original covariances (COV_{ij}), after rounding.

Table 5

Example Three: Varied Item Intercorrelations with Mixed Signs

	Correlation				Variance/Covariance			
Var.	1	2	3	4	1	2	3	4
1	1.00				.22			
2	-1.00	1.00			-.20	.18		
3	-1.00	1.00	1.00		-.20	.18	.18	
4	-1.00	1.00	1.00	1.00	-.18	.16	.16	.15

Note. Covariances were found with $COV_{ij} = r_{ij} (SD_i * SD_j)$, where the standard deviations are the square root of the variance for the variable. Covariances are rounded to two decimal places.

Table 6

Calculation of Total Test Score Variance (σ_{TOTAL}^2) for Example Three

Pairing		COV/Variance			r/SD			
i <	j	COV_{ij}	σ_i^2	σ_j^2	r_{ij}	SD_i	SD_j	COV_{ij}'
1	2	-.20	.22	.18	-1.00	.47	.42	-.20
1	3	-.20	.22	.18	-1.00	.47	.42	-.20
1	4	-.18	.22	.15	-1.00	.47	.39	-.18
2	3	.18	.18	.18	1.00	.42	.42	.18
2	4	.16	.18	.15	1.00	.42	.39	.16
3	4	.16	.18	.15	1.00	.42	.39	.16

ΣCOV_{ij} = -.08
$\Sigma COV_{ij} * 2$ = -.16

Note. COV_{ij}' represents the recalculated covariance using $COV_{ij}' = r_{ij} (SD_i * SD_j)$.

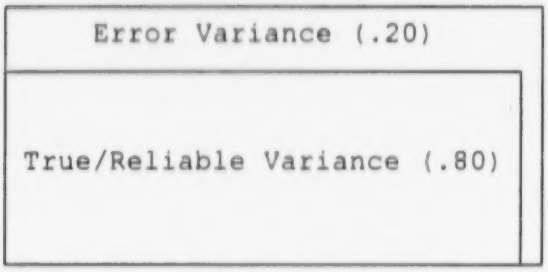

Figure 1. Illustration of classical test theory ratio of true to total score variance (alpha = .80).